DOCUMENT RESUME

ED 211 577 TM 810 953

AUTHOR Jensen, Arthur R.
TITLE Test Validity: g Versus the Specificity Doctrine.
PUB DATE 26 Aug 81
NOTE 71p.; Paper presented at the Annual Meeting of the American Psychological Association (Los Angeles, CA, August, 1981).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Cognitive Tests; *Correlation; Court Litigation; *Intelligence Differences; *Intelligence Quotient; Occupational Tests; *Psychological Testing; *Psychometrics; Scores
IDENTIFIERS *Specificity Doctrine

ABSTRACT

The specificity doctrine, holds that psychometric tests measure nothing other than the specific bits of knowledge and learned skills reflected in the item content of the tests. This prevailing doctrine has influenced the interpretation of test scores and the conceptualization of test validity, as well as the practical use of tests in educational and personnel selection. Opposed to the specificity doctrine is the view that a wide variety of cognitive tests measure in common a few large factors of mental ability, most prominently general intelligence or g. The commonality is not ascribable to common contents of the tests but to common brain processes. Recent massive validity evidence from the use of cognitive tests in personnel selection is consistent with the broad common-factor theory and contradicts the specificity doctrine. The practical and theoretical utility of the construct of g as a general information processing capacity also appears warranted by other lines of evidence independent of factor analysis. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED211577

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Test Validity: g versus the Specificity Doctrine

Arthur R. Jensen

University of California, Berkeley

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A.R. Jensen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ABSTRACT

The specificity doctrine, a legacy of the positivism and radical behaviorism that has dominated the history of American psychology, holds that psychometric tests measure nothing other than the specific bits of knowledge and learned skills reflected in the item content of the tests. This prevailing doctrine has influenced the interpretation of test scores and the conceptualization of test validity, as well as the practical use of tests in educational and personnel selection. Opposed to the specificity doctrine is the view that a wide variety of cognitive tests measure in common a few large factors of mental ability, most prominently general intelligence or g. The commonality is not ascribable to common contents of the tests but to common brain processes. Recent massive validity evidence from the use of cognitive tests in personnel selection is consistent

TM 810 953

with the broad common-factor theory and contradicts the specificity doctrine. The practical and theoretical utility of the construct of g as a general information processing capacity also appears warranted by other lines of evidence independent of factor analaysis.

Test Validity: g versus the Specificity Doctrine

Arthur R. Jensen

University of California, Berkeley

First, I will define the specificity doctrine, and say something about its historical roots and its effects on psychometrics and theories of human abilities. Second, I will describe the manifestations of the specificity doctrine in the applied field of mental testing and personnel selection, and discuss how it has influenced prevailing notions of test validity. Third, I will review the massive research on test validity in the past decade which totally refutes the implications of the specificity doctrine for employment testing, and I will cite estimates of the monetary costs to our nation which would result from the abandonment of tests for employment selection or from the use of less than optimal selection procedures based on test results. Finally, I will point out recent findings from basic research on the nature of g, the general ability factor (less euphemistically known as general intelligence), which may help us to understand theoretically the findings related to the practical validity of selection tests.

The Specificity Doctrine

The essence of the specificity doctrine comprises two beliefs: (1) human mental abilities, and individual differences therein, consist of nothing other than a repertoire of specific items of knowledge and specific skills acquired though learning and experience, and (2) all psychometric tests of mental abilities measure nothing other than some selected sample of the total repertoire of knowledge and skills deemed important by the test constructor.

Various expressions of these beliefs have been seen repeatedly in recent legal decisions regarding the use of tests, in schools and in job selection. Quotations from some court cases vividly illustrate the specificity doctrine.

In the HOBSON v. HANSEN case, which abolished ability grouping in the schools of Washington, D.C., in 1967, Judge J. Skelly Wright stated:

> The skills measured by scholastic aptitude tests are verbal. More precisely, an aptitude test is essentially a test of the student's command of standard English and grammar. The emphasis on these skills is due to the nature of the academic curriculum, which is highly verbal; without such skills a student cannot be successful. Therefore, by measuring the student's present verbal ability the test makes it possible to

estimate the student's likelihood of success in the future. Whether a test is verbal or nonverbal, the skills being measured are not innate or inherited traits. They are learned, acquired through experience. It used to be the prevailing theory that aptitude tests -- or "intelligence" tests as they are often called, although the term is obviously misleading -- do measure some stable, predetermined intellectual process that can be isolated and called intelligence. Today, modern experts in educational testing and psychology have rejected this concept as false. Indeed, the best that can be said about intelligence insofar as testing is concerned is that it is whatever the test measures .... The IQ tests must be recognized as artificial tools to rank individuals according to certain skills.

More recently, in 1979, in the well-known LARRY P. v. WILSON RILES case, which resulted in prohibiting the use of standardized intelligence tests for the placement of black and Hispanic pupils in classes for the mentally retarded in California schools, Judge Robert Peckham, in his final decision, gave succinct expression to the specificity doctrine: "IQ tests, like other ability tests, essentially measure achievement in the skills covered by the examinations .... The tests measure the skills tested, and each of the tests subject to this litigation assesses very similar skills." An even more emphatic statement of the specificity doctrine, almost a caricature of it, by a witness for the plaintiffs

in the Larry P. trial, Princeton psychologist Leon Kamin, is quoted in Judge Peckham's decision, as follows:

> IQ tests measure the degree to which a particular individual who takes the test has experience with a particular piece of information, the particular bits of knowledge, the particular habits and approaches that are tested in these tests.

To which Judge Peckham commented, "This point ought to be an obvious one ..."

Historically, the specificity doctrine is linked to behaviorism and, more generally, to logical positiivism, which was the prevailing philosophy of science for psychology for half a century, from the 1920's to the 1970's (McKeachie, 1976). The highly compatible philosophies of behaviorism and logical positivism viewed scientific psychology as the study of empirically observable behavior, especially experimentally manipulable aspects of behavior (MacKenzie, 1972, 1977). Since sensory inputs could be experimentally manipulated and effector outputs were directly observable as behavior, scientific psychology also tended to be <u>peripheralistic</u>, focusing on variations in stimulus and response elements in its explanations and eschewing reference to more central processes or constructs. The phenomena of conditioning and learning seemed much more compatible with this S-R model for psychology than did the abilities, constructs, and factors which had sprung up in

that branch of psychology that dealt with mental measurement (psychometrics) and the nature of individual differences (differential psychology).

The behaviorist view of mental abilities and mental tests is that abilities are specific, learned elements of behavior, and that mental tests merely measure whether or not a person has learned certain bits of behavior. The way to study intelligence, or any other ability, according to this view, is to observe the specific behaviors demanded by the items of those tests that are termed "intelligence tests," and then determine how and under what conditions all these bits of knowledge and skills are acquired. If there is any organization of abilities, that is, any distinctive pattern of correlations among the various bits of knowledge and skills measured by a test, it is only because certain experiences tend to go together more frequently in the environment than do other experiences. Thus knowledge of symphony and knowledge of opera are more apt to be correlated than either is likely to be correlated with, say, knowledge of baseball.

According to this behavioristic specificity view, mental tests are useful in terms of predictive validity only because they measure specific knowledge and skills that constitute part of the criterion behavior to be predicted by the test scores.

The specificity doctrine ignores or rejects the concept

of broad group factors, especially a general ability factor or general intelligence, because general ability, or g, is not reducible to any specific forms of behavior, knowledge, or skills. Yet the g factor is a phenomenon that cannot be ignored in any complete scientific account of individual differences in ability. A large general factor can be shown to exist in any sizeable collection of complex tests taken by a representative sample of the general population (Eysenck, 1979, 1981; Humphreys, 1982). It makes little difference how the general factor is computed -- as the first principal component, or the first unrotated factor in a common factor analysis, or as a hierarchical second-order factor derived from oblique rotation of the primary factors. Although each of these general factors accounts for a slightly different proportion of the total variance, they are all very highly correlated with one another when a variety of tests is entered into the factor analysis. The g factor is extremely robust across methods of extraction, across batteries of tests, and across different populations (Cooley, 1976; Humphreys, 1981).

It is this concept of general ability, and also perhaps a few broad group factors in addition to g, such as verbal, quantitative, spatial, psychomotor, and perceptual abilities, that opposes the specificity doctrine. The g factor is an insuperable problem for behaviorism, because it simply cannot be described or defined or understood in terms of sheer behavior, or in terms of conditioning and learning. A

great variety of behavioral tasks with no learned elements in common are all capable of measuring the same g. For example, no behavioral analysis of performance on a vocabulary test and a block designs test (as in the Wechsler Intelligence Scales) can account for the high correlation between these diverse tests, which are both good measures of g. The high correlation between such distinctly different batteries, with respect to knowledge, skills, and behavioral elements, as the Verbal and Performance scales of the Wechsler tests absolutely contradicts the specificity doctrine. The specificity doctrine is most strikingly contradicted by the demonstration of the g factor itself, by means of factor analysis, and by the fact that tests as diverse as reading comprehension, backward digit span, and choice reaction time are all loaded on g. There are no simple, specific behaviors, bits of knowledge, or particular skills that are common to many of the tasks that manifest g. The specific contents of intelligence tests merely serve as vehicles for the measurement of g, and many behaviorally diverse vehicles serve equally well. The test content is not the factor. The subject's behavior is not the factor. The medium is not the message. Common elements, if indeed they are the correct explanation for g, are common processes of the brain, not common items of knowledge, skill, or learned behavior.

To be sure, g is a hypothetical construct, the ultimate nature of which is yet to be fully discovered. But let us

hasten to recognize that the very same statement applies to every construct in science -- time, mass, electricity, magnetic field, gravity, and so on. The mechanism of gravity, for example, is still not understood, and there are alternative theories, no one of which successfully comprehends all of the empirical phenomena. Scientists cannot directly observe a magnetic field or gravity, but can only measure their effects on observable objects. In this respect, intelligence is not different from all other scientific constructs. We cannot directly measure intelligence itself, but only its influence on performance. The way a task is performed can indicate the efficiency and capacity of the underlying operating system. As a theoretical construct, g may be thought of as the overall or general efficiency of an information processing system with respect to its capacity and its speed of handling information. This could represent a general level of efficiency of all the subunits of the system -- a more or less homogeneous quality throughout, or it could represent merely an average value of various subunits differing in efficiency. These are theoretical questions that can only be answered by empirical research.

Performance on intelligence tests, and on virtually every other kind of cognitive test, reflects to some extent this hypothetical construct of efficiency of information processing as represented by g. Hence the essential validity of such tests is their construct validity. This is in no way intended to depreciate the broad practical predictive

validity of g-loaded tests for many educationally and occupationally important criteria. But practical predictive validity of such tests, as I will show, cannot be understood in terms of the specificity doctrine. The practical predictive validity of g-loaded tests can only be understood, at present, in terms of the tests' construct validity. But of course the concept of construct validity is itself an anathema to strict adherents of the specificity doctrine, because this doctrine holds that common behavioral elements and common bits of learned knowledge and skills, rather than common factors representing essentially inferred brain processes, are the root of correlations among various tests and of the correlations between tests and job performance.

Group factors, such as verbal, quantitative, spatial, perceptual, and psychomotor abilities, are slightly more compatible with the specificity notion, because, unlike g, they are always tied to particular types of test content and directly teachable knowledge and skills. This fact may account, in part, for the greater popularity, among those who use factor analysis in the study of tests, of Kaiser's varimax and similar methods for orthogonal rotation of factors -- methods that are expressly designed mathematically to submerge the general factor and to reduce the "positive manifold" matrix of intercorrelations among ability tests to a number of orthogonal (i.e., uncorrelated) group factors. This was the aim of Thurstone's method of multiple factor analysis. But the g variance is then merely hidden among

the various group factors, as Thurstone later recognized in his advocacy of oblique rotation of the factors to yield correlated primary factors which could then be factor analyzed to yield a second-order g factor. The really telling point, however, is the fact that it has proved utterly impossible to make up tests that measure the primary or group factors but do not also measure g. In fact, highly "purified" tests of the main group factors -- verbal, numerical, spatial, memory, etc., -- usually are loaded as much or more with g variance as with the variance identified with their particular group factors.

Guilford's Structure of Intellect model of abilities, with its 120 hypothesized narrow orthogonal factors, and no general factor or broad group factors, was much more in tune with the specificity zeitgeist of the 1960's than the hierarchical factor models of Spearman, Burt, and Vernon, with their large general factor and a few broad group factors. But the Structure of Intellect, I think, is already passé; it has proved to be a theoretical blind alley, and also has not proved fruitful in the practical use of tests. It probably represents the most extreme outpost of the specificity doctrine we will ever see in the factor analysis of abilities. In the heyday of the specificity doctrine, Quinn McNemar's (1964) presidential address to the APA, in 1963, entitled "Lost: Our Intelligence. Why?," seemed a lone voice in the wilderness. Today he would have lots of company in psychometrics and differential psychology. The g

factor is again of great interest in the abilities domain, and methods of hierarchical factor analaysis based on oblique rotation, or other methods that highlight g, are now generally preferred over orthogonal rotation. I believe that in the abilities domain it will soon be widely recognized that any method of factor analysis which does not, by its mathematical nature, allow a g factor is psychologically and theoretically inappropriate. Some people today prefer orthogonal rotation for no other reason than that it prevents the appearance of a general factor. At present, the theoretically most attractive method of factor analysis of ability tests is the hierarchical Schmid-Leiman (1957) procedure, which extracts a second-order g factor from a number of oblique primary or group factors and then orthogonalizes the primary factors, so that all factors in the hierarchy are uncorrelated with one another. Not only is it attractive for theoretical reasons that I will not go into here, but it puts the factors and the tests representing them in a proper perspective of importance in terms of the generality of their practical validity. This is because the predictive validity of tests depends upon the degree of congruence between the factor structure of the test and the factor structure of the tasks involved in the job performance or other criterion performance that the test is intended to predict, and the g factor is generally predominant in the criteria, as it is in the tests.

Although recent developments in theory and research on

the psychology of human abilities go counter to the specificity doctrine, the legacy of its 50-years dominance is still prevalent. The specificity doctrine has had a similar paralyzing effect on the science of abilities and on its applied aspects in personnel psychology as vitalistic theories had on chemical and biological research in the 18th and 19th Centuries. The specificity doctrine continues to have great appeal primarily to three groups: (1) those who hold out hopes that it will be possible to raise children's intelligence appreciably by purely behavioral or psychological techniques, because it is easier to see how specific knowledge and skills can be explicitly taught than how to make alterations in an unobservable factor or theoretical construct such as g; (2) those who for ideological reasons would like to diminish the potential importance of tests and their wide use in our society; and (3) those who are looking for easy ways to explain away the observed average racial and ethnic group differences in test performance. The specificity doctrine is also favorable to the notion of the substitutability of abilities and other traits -- the idea that traits such as creativity, special talents, motivation, and the like, can act as substitutes for general intelligence. This notion may seem more comforting, in the face of low g, than the idea of a given level of g being a necessary, albeit not sufficient, condition for success.

Employment Testing and the Specificity Doctrine

The specificity doctrine has made its most costly impact through the agency of federal laws governing the use of tests in employment selection. The EEOC Guidelines on Testing, the Federal Executive Agency Guidlines, and the Uniform Guidelines on Employee Selection Procedures are all solidly entrenched in the specificity doctrine and literally force its various manifestations and consequences on all users of tests in personnel selection (Prien & Ronan, 1971; Schmidt, Hunter, & Pearlman, 1981; Eyde, Prinoff, & Hardt, Note 1; McGuire, Note 2). The specificity doctrine has been a useful weapon to Federal enforcement agencies in their war to eliminate personnel tests -- a war nobly motivated, I presume, by the wish to redress racial imbalances in employment selection associated with observed racial differences in performance. These differences are predictable by means of tests and are brought into sharper focus when tests are used by employers. Racial imbalance in selection arising from objective test predictions of job performance is referred to as "adverse impact." Any employment tests resulting in adverse impact must, according to federal testing guidelines, demonstrate that the tests are specifically relevant to the particular criterion of job performance to be predicted. The courts have placed on the employer the burden of showing that any selection requirement, such as a given score on an aptitude test, must have a "manifest relationship to the employment in question." In a number of

court cases, this has been interpreted to mean that the tests should measure the actual skills needed for the particular job. In terms of the specificity doctrine, criterion-related test validity, that is, the correlation between test scores and job performance, is specific to each particular job, to each employment situation, and to each racial or ethnic subpopulation. According to the guidelines, test validity must be demonstrated by separate validation studies of a given test for every specific combination of every one of these three factors -- jobs, situations, and subpopulations. Specifically job-relevant content validity or specific criterion predictive validity are required of all tests. Although construct validity is ostensibly allowed in the Uniform Guidelines, it is redefined so as to be equivalent to criterion-related validity. Hence factorial validity and construct validity in the broad sense, in effect, are rendered unacceptable. In terms of the Uniform Guidelines, it would be hard for most employers, if adverse impact were claimed, to get away with selection on the basis of an "IQ test" or some other explicit test of general mental ability. In fact, the U.S. Civil Service Commission itself lost a case (DOUGLAS v. HAMPTON) on this very point, by using a general aptitude test (much like the Graudate Record Examination) for the selection of high-level personnel in federal agencies. The Civil Service Commission was not allowed to select employees on the basis of tests of general intelligence and broad verbal and quan-

titative aptitudes, but was required to tailor and validate a number of narrow selection tests, each directly related to the specific performance demands of particular jobs. But in fact such tests were never developed; instead a highly g-loaded test (PACE) was devised for Civil Service employees.

In view of this trend toward specificity in personnel testing, therefore, you can imagine my surprise, when, a couple of years ago, I read the following headline in the news section of Science (June 22, 1979, Vol. 204, p.1285): "IQ Tests for Reactor Operators." The gist of the story can be best conveyed by direct quotes:

> The Tennessee Valley Authority (TVA) is making a bid to become the nation's leader in selection and training of nuclear reactor operators in the wake of the accident at Pennsylvania's Three Mile Island plant. ...the report calls for extending the training period of reactor operators from 2 to 3 years, and for the introduction of "stringent intelligence testing." TVA already routinely gives personality tests to job applicants ... Applicants also take a General Aptitude Test battery that measures mechanical aptitudes. ...would be operators also have to pass the Nuclear Regulatory Commission examination and oral and written TVA examinations. The task force decided that these measures are still inadequate for predicting an operator's performance on the job. From now on, "intelligence will be

> stressed as one of the most important characteristics of superior reactor operators." The report says "intelligence distinguishes those who have merely memorized a series of discrete manual operations from those who can think through a problem and conceptualize solutions based on a fundamental understanding of possible contingencies." General intelligence testing for job applicants -- that is, tests not directly related to job performance -- is very controversial and their use has become highly circumscribed in recent years. But according to a psychologist whom TVA has consulted, "IQ testing is definitely on its way back."

The opinions expressed in this quote remind me of one rather good definition of intelligence as "what you use when you don't know what to do," that is, when the problem isn't something you've already been taught specifically how to solve. This is a characteristic that many employers are seeking in their employees in a wide variety of jobs, and they want it more, the higher and more responsible the job. It should be noted that this is one of the defining features of the construct validity of g, or intelligence. Selection on g, as measured by general intelligence tests, inevitably raises the average level of this highly valued characteristic among those who are employed. Its relative value, of course, will differ among jobs, depending on the complexity of the work, the amount of routine involved, and the closeness of supervision.

Validity of g versus Differential Aptitude Batteries

How does a single highly g-loaded test score compare with a multi-aptitude test battery in its validity for predicting performance in a wide variety of jobs? To get some idea about this, I have tabluated 537 predictive and concurrent validity coefficients obtained in 446 different occupations, based on the General Aptitude Test Battery (GATB) from data provided by the U.S. Employement Service. In Figure 1, I have plotted separately the validity

---
Insert Figure 1 about here
---

coefficients for G, the general intelligence score of the GATB, and the multifactor validity coefficients, based on the (multiple) correlation between an optimally weighted composite of the nine aptitudes of the GATB, including G, for each job. In addition to general intelligence (G), these aptitudes are verbal, numerical, 3-dimensional spatial visulization, form perception, clerical speed and accuracy, motor coordination, finger dexterity, and manual dexterity. The fact that the GATB Aptitude G is a good measure of the g factor, or general intelligence, is shown by its average correlation of .80 with twelve well-known standard tests of IQ and a g-loading of .85 when it is factor analyzed along with all the other GATB aptitudes.

The median G score observed validity is +.27; the median multifactor observed validity is +.36. Before

comparing these validity coefficients further, let us be reminded of the accurate meaning of a validity coefficient. An observed validity coefficient is, of course, the correlation (simple or multiple) between test scores and the criterion -- some quantified assessment of job performance. Like any correlation, the validity coefficient is subject to attenuation from unreliability of measurement in the test and in the criterion. True (or operational) validity is estimated by correcting for these two attenuating factors. The validity coefficient is also subject to attenuation from restriction of variance; and it fluctuates due to sampling error. The proper interpretation of a validity coefficient was explicated long ago by Brogden (1946), who showed that for any given selection ratio, a test's true validity coefficient is a direct measure of the proportional gain in criterion performance resulting from top-down selection of job applicants on the basis of test scores, over the average level of criterion performance that would be obtained if there were random selection of applicants. Thus selection by means of tests with validities in the vicinity of .30 is by no means trivial to an employer -- it represents 30 percent of the gain in the average quality, proficiency, or productivity of the employees' job performance that would result from selection based on a test with perfect validity ($\underline{r}$=1.00).

The difference between the median G observed validity of +.27 and the multifactor observed validity of +.36 is

really more remarkable than may appear at first glance. It is remarkable that the difference -- .09 -- is so small. Notice, first of all, that we are here comparing a simple correlation with a multiple correlation, and a multiple R is always statistically biased upwards.[1] Then, there is also the fact that a multiple R must always be greater than zero, no matter how great the sampling error may be, whereas a simple r can be less than zero due to sampling error even when the true correlation is some positive value. Finally, the multifactor validity is based on a different optimally weighted composite from among the nine GATB aptitudes, including g, for each job category. Also note that the multifactor composite, being based on several tests, necessarily has higher reliability than the single G aptitude. Even if each of the several tests in an optimally weighted composite measured nothing other than the g factor, the validity of the composite would be higher than for any single test. So if these several statistical artifacts were all taken into account, the difference between the multifactor validity and the G validity would probably be reduced to some value scarcely greater than the median G validity of +.27. I point out all this, not to belittle the value of multivariate prediction for many jobs, which cannot be denied, but to show that most of the predictive validity is attributable to g, that is, general intelligence. And this g can be measured with more or less equal reliability and construct validity by a great variety of tests. To estab-

lish that a test is a good measure of g, we need only to observe the size of its g loading when it is factor analyzed among any large battery of diverse cognitive tests. Or the test in question can be shown to be highly correlated with any of a number of standard tests of intelligence, virtually all of which have high factorial validity on g, whatever other group factors they may also measure.

The g factor has predictive validity for job performance for practically all jobs, but the validity of g increases with job complexity -- another fact predictable from the construct validity of g. Jobs requiring the integration and coordination of information have higher g validity than jobs involving compiling or computing data, which in turn have greater g validity than jobs involving checking, comparing, and copying information or executing routine procedures (Hunter, Note 3). For most jobs, g accounts for all of the significantly predicted variance; other testable ability factors, independently of g, add practically nothing to the predictive validity. However, the overall validity for predicting job performance in many of the skilled trades is significantly improved by tests of spatial visualization and psychomotor factors, in addition to g; and the validity of predictions of performance in many clerical occupations is significantly improved by including tests that measure a preceptual speed and accuracy factor, in addition to g. Hunter (Note 3) found that psychomotor abilities became increasingly valid predictors of job

performance as the g-loading of jobs decreased. Even when explicit tests of intelligence are not used, but only selected aptitude tests, such as those of the GATB (except the G aptitude), it should be remembered that some of the aptitude tests are at least as highly loaded on g as on the particular aptitude they are intended to measure, and a good part of their predictive validity is due to their g loading. (Tests of motor ability and dexterity in the GATB have the lowest g loadings, ranging between .20 and .50.) The average predictive validities of each of the GATB aptitude tests, for 300 occupations, are correlated .65 with the g loadings of these aptitude tests (Jensen, 1980a, pp. 735-736). The federal guidelines on testing are clearly on the wrong track, in terms of the economics of testing, by proscribing general ability tests and demanding tests of highly specific job-related content for employment selection.

Prevalent Specificity Myths Debunked

Differential Validity. Until the mid-1970's, it was generally believed that the standard tests used in schools and for college selection and employment testing are not equally valid for majority and minority groups. This belief was bolstered by three influences:

(1) It was consistent with the specificity doctrine, since if it were believed that tests measure only specific bits of information acquired in a particular control context, and that minority groups differ in cultural

background, it should not be surprising that tests could not be equally valid as predictors of job performance for majority and minority persons. It was thought that minority cultural differences would affect test performance much more unfavorably for tests devised and standardized for the majority than they would affect actual job performance. Therefore, it must be required that a test's validity be demonstrated separately for majority and minority applicants. This requirement was written into the early federal guidelines on testing for job selection.

(2) It seemed necessary to social scientists to offer some explanation for the average differences in test scores between certain racial groups, particularly whites and blacks. The most obvious, or at least most socially acceptable, explanation seemed to be that the tests are culture biased in such a way as to favor whites and disfavor blacks and Hispanics.

(3) Quite a number of studies, in fact, suggested that there was single group validity (a significant validity coefficient for one group, majority or minority, but not for the other) and differential validity (significantly different validity coefficients for two or more groups). Such findings were interpreted as supporting the specificity of test validity for groups, and as supporting the cultural bias explanation of the average lower scores of blacks and Hispanics on most tests. The occasional findings of single

group validity and differential validity seemed to justify the demand that every test's validity be established separately for every subpopulation in which it was to be used for employment selection.

The evidence, however, is now overwhelming to the effect that the most widely used standardized tests are not culturally biased against minority groups in the United States (Jensen, 1980a). Hundreds of studies of test bias have shown that tests can be used with equal effectiveness for whites and blacks alike for all the legitimate uses of tests in college selection and employment selection. The same thing can be said for other native-born, English-speaking minorities, including Hispanics. Meta-analyses of all the data reported in the entire research literature in this area support the conclusion that single group validity and differential validity are nonexistent phenomena when comparisons are between whites and blacks (Bartlett, Bobko, Mosier, & Hannon, 1978; Hunter, Schmidt, & Hunter, 1979) or between whites and Hispanics (Schmidt, Pearlman, & Hunter, 1980). Single group validity and differential validity are found in studies occasionally, but with no greater frequency than would be statistically expected from sampling error. The inability to demonstrate statistically the existence of culture bias in widely used tests, given the exceedingly massive data now available, must be interpreted as meaning either that the cultural differences between blacks and whites has been grossly exaggerated by those seeking an easy

explanation for the average white-black difference in test scores, or that our current tests do not interact significantly with whatever cultural differences do exist. I have reviewed the evidence for this conclusion in detail in my book Bias in Mental Testing (Jensen, 1980a) and in a more summary fashion in my most recent book, Straight Talk About Mental Tests (1981). Although my book on Bias has by now received over seventy reviews, not one reviewer has yet challenged its main conclusion, namely, that our current standardized tests are not biased against any of the native-born English-speaking minority groups on which the amount of research evidence is sufficient for an objective determination of bias, if the tests were in fact biased. For most nonverbal standardized tests, this generalization is not limited to English-speaking minorities. The nonexistence of differential validity apparently is already sufficiently recognized by now that the new Uniform Guidelines for testing no longer require separate validity studies for minority groups.

Not only are tests equally valid within both minority and majority groups, but the actual test scores themselves do not underestimate the criterion performance of minority persons -- in college, in training programs, or in jobs.

That is to say, the regression of criterion measures on test scores is the same for all groups. Where blacks or Hispanics score below whites on the selection tests, they

also perform below whites on the job criterion measures, and this is true when the criterion measures are supervisor's ratings, objective work sample tests, job knowledge tests, or indices of actual productivity on the job. In fact, where significant errors of prediction have occasionally been found, it has been that tests have overpredicted minority performance on the criterion (e.g., college grades, job performance). This constitutes bias in favor of the minority group when test scores are used for selection without adjustment for racial group membership. But such overprediction of minority performance often diminishes to nonsignificance when prediction parameters are corrected for attenuation due to the imperfect reliability of the tests. (This is equivalent to using estimated true scores based on each group's own mean and standard deviation.) Another cause of significant group differences in prediction from a common (or majority) regression equation is that certain ability or personality factors that contribute variance to the criterion are not measured by the test (Hunter, Schmidt, & Rauschanberger, in press). When validity is improved by adding the missing variables to the prediction equation, predictive bias with respect to minority and majority groups is greatly reduced or entirely eliminated, making the test equally fair for all applicants regardless of group membership.

It is a significant fact, shown in data from 51 validity studies compiled by Hunter (Note 4), that the degree of

overprediction of black job performance by scores on the nine different aptitudes of the GATB is inversely related to the g loadings of the aptitudes. The rank order correlation between the amount of overprediction (i.e., test bias) and the tests' g loadings is -.80. In other words, the more g-loaded the tests, the less they show predictive bias. This means that job performance criteria in general are themselves substantially g-loaded, and it requires sufficiently g-loaded tests to predict g-loaded job performance with equal accuracy for blacks and whites. This finding absolutely contradicts the specificity notion of unique cultural factors interacting with test items to produce racial-cultural bias in the test scores.

An inevitable and important consequence of the fact that tests are not culture biased and are every bit as valid and fair for minorities as for the white majority is the adverse impact of the use of tests on minority hiring rates, if the optimal top-down or ranking procedure is used for selection. The adverse impact of the optimal selection procedure is greater for those jobs for which the g factor contributes the largest part of the test's predictive validity, because the average difference between whites and blacks is greater on g than on any other factor identified by psychometric tests. This is especially consequential for higher status jobs, which are intrinsically more highly g-loaded.

One might imagine that adverse impact would be largely mitigated by the fact that applicants for different jobs with differing g demands are self-selected and this would diminish the effect of average racial differences on hiring rates. Various jobs do not attract applicants from the total distribution of abilities. Self-selected applicants for each job category come from a particular segment of the ability distribution roughly corresponding to the g demands of the job. An interesting point, however, is revealed by the data on 38,452 black and 142,545 white self-selected job applicants in 80 different occupations who have taken the Wonderlic Personnel Test (WPT), a measure of general intelligence (Wonderlic, 1972). There is a very high correlation (.87) between the black average and white average WPT scores over the 80 job categories, indicating that the same sort of self-selection for jobs operates within both racial groups. This, however, does not result in smaller average differences between black and white applicants within job categories, when the differences are expressed in terms of the standard deviation within each job category. The reason for this is apparent in the Wonderlic data: job applicants are self-selected, not according to an absolute standard of aptitude requirements for each job category, but according to their relative positions in their own racial group's distribution of aptitude. Thus, if white applicants in a particular job category have an average raw score which is at, say, the 60th percentile of the total white score distribu-

tion for Wonderlic national norms, the average raw score of black applicants in the same job category will be at approximately the 60th percentile of the total black score distribution. But a test score at the 60th percentile of the total score distribution for blacks represents a lower absolute level of ability, by about one standard deviation, than is represented by a test score at the 60th percentile for whites. Because the variation of applicants' scores within a given job category is usually less than the variation in the total population, the average white-black difference (in standard deviation units) is often as great or greater than the one standard deviation difference between the black and white populations in general. Thus self-selection of applicants in the job market in no way overcomes the problem of adverse impact, and, if anything, may even magnify it.

All proposed remedies for adverse impact necessarily involve some form of less than optimal selection procedure, if optimal is defined as that procedure which maximizes the quality and productivity of the group selected for employment. The optimal procedure, assuming test scores are unbiased, consists of a totally colorblind top-down selection or ranking of the applicants. Any selection procedure other than colorblind top-down selection amounts to some form of quota. It is bound to be less efficient and therefore more costly to the employer. Naturally, the costs are passed on to the consumer and ultimately to the general public in the form of increments to inflation and reduced

competitiveness with other equally industrialized nations in the world market. There is a necessary trade-off between attempts to achieve racial balance and higher labor costs. To what extent employers and society in general can afford less than the optimal selection procedures that the best testing practices can easily make possible today is not a scientific question, but a problem of social philosophy and economic realities. Debates about policy decisions in this area highlight two basically conflicting philosophies governing selection strategy: an individualist philosophy that emphasizes equality of opportunity to all individuals, regardless of race, sex, or national origin, versus a collectivist philosophy that emphasizes immediate equality of representation for various groups in the population. These two philosophies dictate selection strategies that, given the reality of equal test validity for minority and majority groups, are intrinsically incompatible and in direct opposition.

If anything less than an optimal selection procedure is practiced at entry level jobs in order to minimize adverse impact on minorities, the adverse impact is merely postponed to a later time, when employees come up for promotions within the organization. Then promotions to more g-demanding higher-level jobs are likely to create racial imbalances even if promotion is based only on past work performance, without the use of standard tests in deciding promotions.

Two main methods have been used to reduce adverse impact. One method, unfortunately the most popular, is to set a very low cutting score for selection and then hire at random from the pool of applicants who score above the cut-off. The usual justification given for this strategy is what I term the ability threshold fallacy, which holds that above some rather minimal threshold level of ability needed to perform in a particular job, individual differences in tested ability are no longer reflected in performance. This is a fallacy because, for virtually all types of performance the regression of job performance on tested general ability is linear throughout the full range of ability. Restriction of range-of-talent and low criterion reliability, in addition to other statistical and selection artifacts, are largely responsible for the results of many small-scale studies that give a contrary impression. Schmidt and Hunter (1980) have pointed out that avoiding adverse impact by setting low minimum standards for all employees is the most inefficient and costly method of selection. It is almost as bad as having no selection standards at all. Moreover, it does relatively little to reduce adverse impact.

A much better method for reducing adverse impact is top-down selection, based on test scores, separately within the minority and majority groups, to achieve whatever degree of racial balance is explicitly desired. This is quota selection, of course, but top-down selection within groups insures the least costly racial balance, in terms of work

productivity, that can be achieved. There can be an additional hidden cost, however, as noted by Schmidt and Hunter (1980). I call it the performance diffusion effect. It is the insidious lowering of performance standards of nearly all employees as a result of admitting a special class of employees who are underqualified and for whom lower standards of performance are tolerated. Their lower standards of performance diffuse throughout the organization's workforce, producing a greater effect in lowering the overall work standards than could be attributed to just the direct effects of the small group of underqualified selectees. Adequate studies of the full impact of the diffusion effect in different types of organizations are called for.

Validity Generalization. The conclusions reported in this section are derived entirely from the brilliant researches of John E. Hunter and Frank L. Schmidt and their co-workers (see references). During the 1970's, Hunter and Schmidt pursued a program of research that has cogently disproved all the most popularly believed myths about the manifestations of the specificity doctrine in the field of personnel testing and selection. In my opinion it is the most outstanding research of the decade in this field. Essentially, Hunter and Schmidt have developed a method of meta-analysis which they have systematically applied to the statistical results of hundreds of independent studies of test validity employing hundreds of thousands of subjects. (Meta-analysis is a set of statistical methods for

integrating and properly interpreting the overall results of research findings across numerous independent studies [Glass, 1976].) This meta-analysis compellingly demonstrates that most of the variation in validity coefficients across different studies, different jobs, and different situations is attributable to a host of statistical artifacts, such as sampling error and variation in criterion reliability, test reliability, range restriction, as well as criterion contamination, computational and clerical errors, and differences in the factor structures between tests of a nominally given type. Most of the interactions invented to account for differences in the validity coefficients in different studies, according to Hunter, Schmidt, and Jackson (Note 5), "are nonexistent, apparitions composed of the ectoplasm of sampling error and other artifacts."

These meta-analytic studies have dispelled the myth of the <u>situational specificity</u> of employment tests -- the notion that a test that is a valid predictor for a given job in some companies is invalid in other employment situations, and that a separate validity study is therefore needed by each company using the test. This belief is not only generally false for most tests, it is especially false with respect to tests of general ability, which are valid for <u>all</u> supervisory and higher-level jobs involving complex judgments and flexibility, regardless of the specific setting in which the jobs are performed.

Another important conclusion revealed by Hunter's and Schmidt's meta-analysis of hundreds of validity studies is that validity coefficients for tests with substantial g loadings are not job-specific. Hence, for general ability tests, validity does not have to be determined anew for each job for which the test is used for selection. The validity coefficients of most widely used personnel selection tests are generalizable across jobs classified into at most a few broad job categories, such as (a) unskilled labor, (b) skilled manual jobs, (c) clerical jobs, and (d) supervisory, managerial, technical and professional jobs. The validity of cognitive tests is significant in all job categories, but increases going from unskilled to managerial and professional. Other factors besides g will enhance validity in certain categories, for example, spatial and psychomotor factors in skilled manual jobs and perceptual speed and accuracy in clerical jobs. But g is the "basic predictor," to use Hunter's (Note 3) term. Meta-analysis reveals that practically all of the variation in general aptitude validity coefficients across jobs found in numerous studies is the result of sampling error and other statistical artifacts (Schmidt & Hunter, 1978; Schmidt, Hunter, Pearlman, & Shane, 1979). True differences in test validity, even among entirely different jobs, are very small (Schmidt, Hunter, & Pearlman, 1981). The fact that the moderating effect of different work tasks on general aptitude tests' predictive validity is practically negligible, even when jobs differ

grossly in task makeup, clearly contradicts the specificity rationale for the test validation procedures required by the Uniform Guidelines. Fine-grained job analyses or task analyses as a basis for devising selection test batteries are unnecessary. To take an example that would be familiar to most psychologists, imagine that we wanted to predict people's performance on the Verbal Scale of the Wechsler Adult Intelligence Scale (WAIS) and also on the Performance scale, using some other predictor tests. A fine-grained task analysis of the Verbal subtests and of the Performance subtests might easily lead one to expect that quite different kinds of tests would be needed to predict scores on the Verbal and Performance scales of the WAIS. Yet a single highly g-loaded test would predict both the Verbal and Performance scales about equally well, with a validity above .80. To the task analyst, it would most likely come as a surprise to discover that the correlation between scores on the WAIS Verbal and Performance scales, with their apparently very different task demands, is about +.80 in the standardization population.

The reason for the broad generalizability of the validity of cognitive tests across so many jobs and situations is not that there is a dearth of specificity variance in jobs, in addition to broad ability factors. Specificity variance is indeed extremely plentiful. It contributes to the rather moderate ceiling, between .5 and .7, for test validity. But the prospect of devising tests whose

cognitive specificity variance matches the specificity of any particular job is unfeasible and perhaps even impossible. The specific "factors" in cognitive tests, left over after g and two or three large group factors are extracted, are inconsequential contributors to test validity. Specificity, in both tests and jobs, consists mainly of reliable but more or less unique subject-by-task interactions, and is therefore necessarily unable to serve as a predictor variable for a number of individuals. On the other hand, we can say for sure that to the extent that any job requires thinking, any g-loaded test will be a valid predictor of job performance. The prediction ceiling for cognitive tests also results from the likelihood that they do not reflect certain personality and motivational variables that enter into certain types of job performance. For example, the repeated finding of lower than average validities for cognitive tests for sales jobs suggests that personality factors are important in these jobs.

Another prevalent belief in the specificity vein is that tests that are valid predictors of success in training are often not valid predictors of performance on the job. But meta-analysis again undermines this belief (Pearlman, Schmidt, & Hunter, 1980). It turns out that cognitive tests rank order subjects in about the same way with respect to both job training and job proficiency. Tests that are valid for one are valid for the other. Meta-analysis of numerous studies has clearly shown that the same abilities predict

both training success and job proficiency. When selection tests are categorized into five broad test types, the correlation between average training criterion validities and job proficiency validities for a variety of jobs is around .80 (Pearlman, Schmidt, & Hunter, 1980; Lilienthal & Pearlman, in press). However, to the extent that tests are g-loaded, they yield somewhat higher validities for training criteria than for job proficiency criteria. This finding, too, is expected in terms of what we know about the construct validity of g.

The Economics of Testing

If the specificity notions of test validity that prevail through federal enforcement agencies lead to unnecessary and costly validity studies, or to greatly relaxed selection standards, or to quotas, or to abandoning the use of tests altogether in an organization, it would seem important to estimate the actual monetary consequences of these alternatives. If tests are abandoned or replaced by less valid selection techniques, such as interviews, there is bound to be a decline in the overall quality of those hired; and lower-performing personnel means a loss in productivity. Relaxed selection standards greatly increase the costs of training programs, and these costs, along with those of enforced unnecessary validation studies, are passed on to consumers in the form of higher prices. These higher prices, in turn, reduce buying, decrease competition in

international markets, and contribute to unemployment and inflation. If obstacles to the use of the available optimal selection procedures are sufficiently widespread, the result could seriously affect the gross national product. According to TIME, the rate of productivity growth in the U.S. has declined, since 1970, from 3.5% to 1%. Schmidt (1979) suggests that this decline is due in part to a reduced efficiency in allocating people to jobs, because of governmental obstacles put in the way of using tests and optimal selection models.

But what is the actual dollar magnitude of selection with and without tests? One indication: a few years ago it was estimated that the total annual savings in training costs to the armed forces as a result of test selection and classification of enlisted personnel was $442 million. The relevance of scores on general ability tests in armed forces training programs is shown, for example, by the fact that recruits in the lower third of the distribution of general ability test scores require two to four times as much training time as recruits in the upper third to attain satisfactory levels of performance in such relatively simple skills as rifle assembly, missile preparation, and map plotting. There is no reason to believe the situation is materially different for civilian employees in our nation's industries.

Hunter and Schmidt (in press) have improved on earlier statistical methods for such cost estimates of the conse-

quences of different selection strategies for various jobs. Their analysis is a real eye-opener, because it turns out that the probable costs of ignoring the kinds of information that can be provided by tests for job selection in all spheres of our national economy are indeed very much greater than most of us might imagine. Hunter's and Schmidt's conclusions are based essentially on statistical estimates of the standard deviation of the dollar value of job performance among randomly selected applicants. Estimates of this standard deviation for various jobs fall between 40% to 70% of the average salary for the job. Employee differences in job proficiency, therefore, correspond to considerable differences in the actual dollar value of their performance. The use of valid tests and optimal selection procedures, of course, simultaneously narrows the range of these differences and raises the average level of job proficiency. This is the basis for estimating the cost effectiveness of tests and selection strategies.

Here are some examples of Hunter's and Schmidt's estimates:

In a study of budget analysts, Schmidt and Hunter (1980) estimated that the dollar value productivity of superior performers (top 15 percent) was $23,000 per year greater than that of low performers (bottom 15 percent). Computer programmers showed a comparable difference. Hunter and Schmidt point out that when these dollar losses are mul-

tiplied by the number of employees in an organization and by the number of years they are employed, the losses quickly mount into millions of dollars.

In a study of the Philadelphia Police Department, with 5000 employees, Hunter (Note 6) estimated that abandonment of a general ability test for the selection of police officers would cost a total of 180 million dollars over a ten year period.

The estimated gain in productivity resulting from one year's use of a more valid selection procedure for computer programmers in the Federal government range from $5.6 to $92.2 million for different sets of estimation parameters (Hunter and Schmidt, Note 7). For the whole Federal government, with 4 million employees, Hunter and Schmidt conservatively estimate that optimal selection procedures would save $16 billion per year.

Hunter and Schmidt (in press) have also estimated the cost effectiveness of using tests for job selection on a national scale. They estimate, for example, that the difference in yearly productivity between random assignment of the workforce to jobs and assignment based on a test with an average true validity of only .45, applied in a working population of 90 million, would be about $169 billion. If general ability tests, to the extent that they are currently used in selection, were to be abandoned, the estimated loss in national productivity would be about $80 billion per

year. If current selection standards were relaxed overall to amount to a selection cut-off at the 33rd percentile of the distribution of test scores, with the top two-thirds of the total distribution being selected, there would be an estimated productivity loss of $54 to $60 billion. On the other hand, if more optimal test selection procedures were practiced throughout the entire economy, Hunter and Schmidt estimate that the GNP would be increased by $80 to $100 billion per year. Even conceding the possibility of a fairly wide margin of error in these estimates, it is apparent that the economic consequences of using selection tests are far from trivial. Their use ultimately affects the nation's overall standard of living. The level of general ability of a nation's population is a natural resource far more important than its physical resources for the quality of life of its citizens. R.B. Cattell and H.J. Butcher (1968) put it well in their book, The Prediction of Achievement and Creativity: "The standard of living of a country is, in the end, not dependent on visible natural resources, or monetary tricks of the economist, but is a function of the level of attainment and creativity prevailing among its citizens" (p.v). Appropriate tests, properly used, can help to identify, cultivate, and deploy a nation's natural resource of ability, to the benefit of individuals and the whole society.

Advances in the Construct Validity of g

Finally, I will mention briefly a few recent lines of investigation that may lead to further understanding of the general ability factor which figures so prominently in the practical predictive validity of tests for so many kinds of performance criteria. The predictive validity of a test, of course, is a result of the degree of congruence between the factor structure of the test and the criterion. The predominant source of predictive variance, as we have seen, is g. This naturally means that the criterion job performance tasks, however different from the tests they may appear to be, are also g-loaded, not because there are common elements of knowledge and skills between the tests and criterion, but because there are common processes at some deeper level of brain functioning.

Cognitive Complexity and g. The core of the construct validity of g is best characterized, for want of a better terminology, as complexity of information processing. As Cattell and Butcher (1968) have said, "The limit of a person's intelligence ... depends upon the degree of complication in the relations that he can perceive, regardless of what fundamentals the relation deals with" (p. 17). We are now discovering that the relationship of g to task complexity is manifested throughout the range of task complexity, from slight differences in complexity in tasks of the least complexity imaginable, all the way up to the most complex

cognitive test items ever devised.

Years ago, it was noticed by Spearman (1927) and Hull (1928) that the more complex cognitive tests showed higher intercorrelations among one another and higher correlations with all other tests, and consequently had higher g loadings, than did seemingly less complex perceptual and motor tests, which displayed relatively little variance in common among themselves or with other tests. But it was not entirely clear whether this finding was due to differences between tests in the sensory and motor modalities or differences in item content. So now we have tried to look at the relationship of g loadings to differences in task complexity among tasks that are very homogeneous in terms of test format, content, and fundamental task requirements. It was our study of forward and backward digit span (in the Wechsler Digit Span subtest) that first got me onto this tack. The backward digit span task, which presumably involves somewhat more complex mental operations than forward digit span, has about double the g loading of forward digit span (Jensen & Figueroa, 1975; Jensen & Osborne, in press). This relationship holds up among blacks and whites, ages 5 to 13, and in all socioeconomic levels. Also, the average white-black difference on backward digit span is about double that on forward digit span -- a difference that is not attributable to the average white-black difference in socioeconomic status.

Reaction Time and g. The reaction time (RT) paradigm that I have been researching recently permits variation in task complexity at an exceedingly simple level (Jensen, 1980b; 1982a, 1982b; Jensen, Schafer, & Crinella, 1981; Vernon, Note 8). The essential task consists of no more than removing one's index finger from a pushbutton as quickly as one can when a light, located six inches above the pushbutton, goes on. The onset of the light is preceded by an auditory ready signal ("beep") at random intervals of 1 to 4 seconds. Complexity is manipulated by varying the number of light bulbs displayed, one of which, at random on each trial, is the reaction stimulus. Increasing the number of bulbs in the array thus increases the amount of uncertainty as to which light will go on following the "beep." RT increases as a function of the amount of uncertainty in the display. This increase in RT is a perfectly linear function of the amount of uncertainty when it is expressed as bits of information. (One, two, four, and eight light alternatives correspond to 0, 1, 2, and 3 bits.)

Measurements of various parameters of this paradigm (intercept and slope of the regression of RT on bits, and intra-individual variability, i.e., the standard deviation of RT from trial-to-trial) are found to be correlated with psychometric measures of general mental ability. Such RT-derived measurements, combined in a multiple regression equation, predict some 50 percent or more of the variance in IQ, or g. They are also correlated with evoked electrical

potentials in the brain, the amplitude and latency of which are also related to psychometric g (Jensen, Schafer, & Crinella, 1981). Interestingly, the one RT parameter most highly related to g is intraindividual variability, that is, the consistency of the individual's RT from trial to trial. This finding suggests a "neural oscillation" model as the physiological basis of g, a more rapid rate of oscillation between refractory and excitatory phases of the neural processes resulting in a faster rate of information processing. It is a reasonable hypothesis that individuals differ in the amounts of knowledge and skills called for by ordinary IQ tests, in part, because they differ in the rates with which they process the information offered by the environment. Other things being equal, individuals with greater speed of information processing acquire more cognitively integrated knowledge and skill per unit of time that they interact with the environment. Seemingly small individual differences in speed of information processing, amounting to only a few milliseconds per bit of information, when multiplied by months or years of interaction with the environment, can account in part for the relatively large differences observed between individuals in vocabulary, general information, and the other developed cognitive skills assessed by IQ tests.

But regardless of what the ultimately correct theory of the relationship between RT and psychometric g will consist of, the important point at this stage of our research is

that the RT paradigm is so extremely simple as to involve virtually no demands on past learning or experience, or on any developed skills, or on memory. But it involves variations in complexity of information processing at the lowest possible levels of informational complexity. In terms of traditional task analysis, this RT paradigm has practically nothing in common with the rather wide variety of highly g-loaded psychometric tests with which it has shown substantial correlations. I regard this finding as a direct disproof of the specificity hypothesis. The relationship of IQ or g to RT parameters, and to the latency and amplitude of evoked potentials, can only mean that our standard IQ tests tap fundamental processes involved in individual differences in intellectual ability and not merely differences in specific knowledge, acquired skills, or cultural background.

Moreover, the RT paradigm further substantiates the core construct validity of g. In three independent studies the negative Pearsonian correlations between RT and psychometric g were found to increase in magnitude (absolute value) as a linear function of bits of information in the reaction stimulus array, as shown in Figure 2, from a study by Lally and Nettelbeck, 1977.

Insert Figure 2 about here

Inbreeding Depression and g. Inbreeding depression is a genetic phenomenon manifested in the offspring of parents

who are genetically more closely related than the average degree of relationship among parents who are paired at random in the population. Matings between brother and sister, between father and daughter, or between (first or second) cousins are all examples of inbreeding. The offspring of such matings are said to be inbred. Such inbred offspring show a depression or diminution in those characteristics, physical or mental, which are in some degree genetically influenced by directional dominance. Directional dominance refers to the consistent quantitative alteration of a characteristic, either to increase its observable expression (positive directional dominance) or to decrease it (negative directional dominance), as the result of the influence of dominant and recessive alleles at a given gene locus on a chromosome. Alleles are alternative forms of a gene, which can be symbolized as A and a. If the quantitative effect of A is to increase the trait and of a is to decrease it, and if the phenotypic (i.e., observable effect of the combination Aa at a given locus is not exactly intermediate between the phenotypic effects of the combinations AA and aa, the theoretical explanation is that one of the alleles is dominant and one is recessive. When there is complete dominance, the phenotypic effect of Aa equals the phenotypic effect of AA, if A is dominant and a is recessive. Because A quantitatively enhances the phenotypic trait, the allelic combination Aa is said to represent positive directional dominance. (Because a quantitatively diminishes the pheno-

typic trait, if a were dominant and A were recessive, the effect of the combination Aa, with complete dominance, would be equal to the combination aa, and would represent negative directional dominance.) When dominant alleles enhance a trait and recessive alleles at the same gene loci detract from the positive expression of the trait, inbreeding, i.e., the mating of genetically related parents, increases the chances that recessive alleles from each parent will be paired at the same loci on the chromosomes, thereby diminishing the phenotypic expression of the trait, as compared with its expression if two dominant alleles (e.g., AA) or a dominant allele and a recessive allele (e.g., Aa) were paired at those loci. This phenotypic result is termed inbreeding depression. The degree of inbreeding depression for any given quantitative trait is assessed by comparing measurements of the trait in the inbred offspring of genetically related parents with measurements in the offspring of unrelated parents who are matched with the related parents in trait-relevant characteristics. Inbreeding depression is manifested in many quatitative human characteristics: birth weight, height, head circumference, chest girth, muscular strength, fetal and infant viability, resistance to infectious disease and dental caries, rate of physical maturation, age of walking, and certain mental abilities, including general intelligence. I have explicated the genetical theory of inbreeding depression in greater detail elsewhere (Jensen, 1978).

Before explicating the relevance of inbreeding depression to g, it must be pointed out that the amount of directional dominance involved in the genetic inheritance of a trait is positively related to directional selection for the trait, either by artificial selection or by natural selection in the course of evolution. Selection results from factors in the environment that have differential effects on the survival or reproductive rate of individuals. Positive directional selection refers to a higher rate of reproduction of viable offsprings that possess the trait-enhancing allele. As selection acts upon phenotypes, the population variance attributable to the purely additive effects of genes, which are directly expressed in the phenotype, is decreased at a much faster rate than the variance due to nonadditive genetic effects such as dominance, in which certain phenotypic expressions of different alleles occur only when they are paired in particular combinations. Selection cannot act upon recessive traits in "carriers;" it can only act on the phenotypic expression of the recessive allele if it is paired with another recessive allele (e.g., aa) and is thereby expressed in the phenotype, on which selective factors can act. Hence directional selection for a trait over many generations increases the proportion of individual differences variance in the trait attributable to genetic dominance. Traits which in a Darwinian sense have had a selective advantage in the course of evolution will have accumulated more dominance, and to that extent they will

also exhibit greater inbreeding depression.

If g is a Darwinian fitness character enhancing man's capability for survival, it would have been subjected to selection over many generations and would have accumulated considerable dominance. Therefore, one should expect to find inbreeding depression on IQ tests or other highly g-loaded tests. Moreover, across various tests the amounts of inbreeding depression should be positively correlated with the g loadings of the tests. This correlation need not be perfect, because other ability factors might also be subject to inbreeding depression. But if the g factor were no more than a figment of mathematical machination, it would be a far-fetched and highly improbable prediction that the g loadings of various tests should be directly related to the degree of inbreeding depression manifested on the tests. On the other hand, the finding of a correlation between g and inbreeding depression would suggest that g reflects some biological reality, a fitness trait fashioned through natural selection in the course of human evolution.

I have sought appropriate data which might test this hypothesis. The best available data is from a study of inbreeding depression in Japan, based on a comparison of the offspring of first and second cousin marriages with the offspring of unrelated parents (Schull & Neel, 1956). The two large samples were statistically equated on a number of background characteristics related to IQ. The Japanese ver-

sion of the Wechsler Intelligence Scale for Children was obtained on all subjects, in addition to a host of physical measurements. A number of the physical traits showed inbreeding depression, and so did the WISC IQ. The average degree of inbreeding depression for IQ in the offspring of first and second cousins is 3 to 4 IQ points. In terms of standard deviation units, the amount of IQ depression was comparable to those physical traits that showed the largest effects of inbreeding, such as stature.

To test the hypothesis under coonsideration, I have calculated the g loadings (first principal component) of the eleven WISC subtests used in this study. The relation of the g loadings to the amount of inbreeding depression is shown in Figure 3. The two variables are significantly

---

Insert Figure 3 about here

---

correlated (r=+0.76, df=10, p<.01). It was statistically determined that the slightly differing reliabilities of the various WISC subtests are not in the least responsible for the correlation between g loadings and inbreeding depression on the subtests. Thus the hypothesis seems to be well borne out by these data. The only other possibility, which does not seem great, for an explanation of these results in terms other than inbreeding depression, would be that socioeconomic differences are related both to inbreeding and to g, and the SES index used to equate the inbred and con-

trol samples statistically on SES did not pick up all of the g-related SES variance. (A detailed account of this study is presented elsewhere [Jensen, Note 9].)

The Spearman Hypothesis of the White-Black Difference

It was noted earlier that "adverse impact" resulting from job selection based on test scores was greater for blacks when selection was for the more complex higher-level jobs and when it was based on more highly g-loaded tests. This outcome was presaged by Spearman in his famous work The Abilities of Man (1927). He put forth a conjecture, based on a small amount of evidence, that I have termed the Spearman hypothesis (Jensen, Note 10). Spearman (1927, p.379) conjectured that the varying magnitudes of the mean differences between whites and blacks in standardized scores on a variety of mental tests are directly related to the size of the tests' loadings on g.

I have examined this hypothesis in nine independent studies in which a battery of diverse tests was administered to large white and black samples. I have extracted a g factor (first principal component) from each battery for the combined racial groups. (In each study there was a very high congruence between the g factor loadings of both races.) The g factor loadings, corrected for attenuation when possible, were then correlated with the size of the mean white-black difference expressed in standard score units. In every one of the ten studies, Spearman's

hypothesis is borne out. I have found no study that contradicts it. The tests' g loadings are significantly, and usually highly, correlated with the magnitude of the mean white-black difference. No other factor revealed by these factor analyses shows anywhere near the same degree of relationship to black-white differences as the g factor. So small are the contributions of other factors to the white-black difference that it is not yet entirely clear that they make any significant contribution at all, although this possibility cannot yet be ruled out.

The largest set of relevant data is based on the GATB, with black and white Ns of well over one thousand. Figure 4 shows the g loadings, the white-black mean difference (in

---

Insert Figure 4 about here

---

standard score units), and the average correlation of each aptitude scale with 12 well-known standard tests of IQ, or general intelligence, for each of the nine aptitudes measured by the GATB. Note how closely the profiles for these three variables parallel one another -- a striking substantiation of Spearman's hypothesis. An index of the degree of similarity among the three profiles is the correlation (Pearson r) between profiles. The profile intercorrelations after correction for attenuation (using the profile reliabilities) are shown in parentheses.

g loading x correlation with IQ:	r = .95(.98)
g loading x white-black difference:	r = .88(.90)
Correlation with IQ x white-black difference:	r = .98(1.00)

These results absolutely contradict the implications of the specificity doctrine for understanding the nature of the white-black differences in psychometric tests. These GATB data and all the other evidence I have been able to find that is relevant to Spearman's hypothesis indicate that the black-white difference in test scores is not mainly attributable to idiosyncratic cultural or linguistic peculiarities in this or that test, but to g, the general factor which all sorts of mental tests measure in common but which some tests measure to a greater degree than others.

Finally, let me say that I am fully aware that many of the findings and interpretations presented throughout this paper will be viewed as highly controversial and even as entirely unacceptable by many people. Although I have considerable confidence in all of the empirical findings I have reported, I am much less confident about any views as to their social meaning and their implications for public policy. I have avoided offering my own opinions in this domain. I do hope, however, that the purely empirical and theoretical elements of this total picture will be subjected to the crucible of critical scrutiny. If they are ultimately destined to be discredited, I hope it will be for

truly scientific reasons, rather than because of political, ideological, or sentimental prejudices masquerading as scientific criticism. Here the words of Bertrand Russell seem most apt: "Ethical considerations can only legitimately appear when the truth has been ascertained; they can and should appear as determining our feelings toward the truth, and our manner of ordering our lives in view of the truth, but not as themselves dictating what the truth is to be."

Reference Notes

1. Eyde, L.D., Primoff, E.S., & Hardt, R.H. What should the content of content validity be? Paper presented at the meeting of the American Psychological Association, New York, September 1979.

2. McGuire, J. Testing standards and the legacy of behaviorism. Unpublished manuscript. George Washington University, 1979. (Available from J. McGuire, Arlington County Personnel Department, 2100 N.14th St., Arlington, Virginia 22201.)

3. Hunter, J.E. The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance. Unpublished manuscript, September 17, 1980.

4. Hunter, J.E. Fairness of the General Aptitude Test Battery (GATB): Ability differences and their impact on minority hiring rates. Unpublished manuscript, March 16, 1981.

5. Hunter, J.E., Schmidt, F.L., & Jackson, G.B. Integrating research findings across studies. Unpublished manuscript, May, 1981.

6. Hunter, J.E. An analysis of validity, differential validity, test fairness, and utility for the Phi-

ladelphia Police Officers Selection Examination prepared by the Educational Testing Service. Report to the Philadelphia Federal District Court, Alverez vs. City of Philadelphia, 1979.

7. Hunter, J.E., & Schmidt, F.L. Noncompensatory aspects of the utility of valid personnel selection. Unpublished manuscript, 1980.

8. Vernon, P.A. Speed of information processing and general intellience. Unpublished Ph.D. dissertation. University of California, Berkeley, 1981.

9. Jensen, A.R. Effects of inbreeding on mental ability factors. Unpublished paper, January, 1982.

10. Jensen, A.R. The nature of the averge difference between whites and blacks on psychometric tests: Spearman's hypothesis. Article rejected by American Psychologist, 1979. (Copies available, along with rejection letter from AP and referees' comments, on request.)

References

Bartlett, C.J., Bobko, P., Mosier, S.B., & Hannan, R. Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 1978, 31, 233-241.

Brogden, H.E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 1946, 37, 65-76.

Cattell, R.B., & Butcher, H.J. The prediction of achievement and creativity. New York: Bobbs-Merril, 1968.

Cooley, W.W. Who needs general intelligence? In L.B. Resnick (Ed.), The nature of intelligence. Hillsdale, N.J.: Erlbaum, 1976. Pp. 57-60.

Eysenck, H.J. The structure and measurement of intelligence. London: Springer, 1979.

Eysenck, H.J. The nature of intelligence. In M.P. Friedman, J.P. Das, & N. O'Connor (Eds.), Intelligence and learning. New York: Plenum, 1981. Pp. 67-85.

Glass, G.V. Primary, secondary, and meta-analysis of research. Educational Research, 1976, 5, 3-8.

Guilford, J.P. Three faces of intellect. American Psychologist, 1959, 14, 469-479.

Hull, C.L. Aptitude testing. New York: World Book Co., 1928.

Humphreys, L.G. The primary mental ability. In M.P. Friedman, J.P. Das, & N. O'Connor (Eds.), Intelligence and learning. New York: Plenum, 1981. Pp. 87-102.

Humphreys, L.G. General intelligence. In C.R. Reynolds & R.T. Brown (Eds.), Perspectives on bias in mental testing. New York: Plenum, 1982.

Hunter, J.E. Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB). Washington, D.C.: U.S. Employment Service, U.S. Department of Labor,1980.

Hunter, J.E., Schmidt, F.L., & Hunter, R. Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 1979, 86, 721-735.

Hunter, J.E., Schmidt, F.L., & Rouschenberger, J. Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C.R. Reynolds & R.T. Brown (Eds.), Persepctives on bias in mental testing. New York: Plenum, in press.

Jensen, A.R. Genetic and behavioral effects of nonrandom mating. In R.T. Osborne, C.E. Noble, & N. Weyl. (Eds.)

Human Variation: Psychology of age, race, and sex. New York: Academic Press, 1978. Pp. 51-105.

Jensen, A.R. Bias in mental testing. New York: The Free Press, 1980. a

Jensen, A.R. Chronometric analysis of mental ability. Journal of Social and Biological Structures, 1980, 3, 103-122. b

Jensen, A.R. Straight talk about mental tests. New York: Free Press, 1981.

Jensen, A.R. Reaction time and psychometric g. In H.J. Eysenck (Ed.), A model for intelligence. New York: Springer, 1982. a

Jensen, A.R. The chronometry of intelligence. In R.J. Sternberg (Ed.), Recent advances in research on intelligence. Hillsdale, New Jersey: Erlbaum, 1982. b

Jensen, A.R., & Figueroa, R.A. Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. Journal of Educational Psychology, 1975, 67, 882-893.

Jensen, A.R., & Osborne, R.T. Forward and backward digit span interaction with race and IQ: A longitudinal developmental comparison. Indian Journal of Psychology, in press.

Jensen, A.R., Schafer, E.W.P., & Crinella, F.M. Reaction time, evoked brain potentials, and psychometric g in the severely retarded. Intelligence, 1981, 5, 179-197.

Lally, M., & Nettelbeck, T. Intelligence, reaction time, and inspection time. American Journal of Mental Deficiency, 1977, 82, 273-281.

Lilienthal, R.A., & Pearlman, K. The validity of federal selection tests for aid/technicians in the health, science, and engineering fields. Washington, D.C.: U.S. Office of Personnel Management, Personnel Research and Development Center, in press.

Mackenzie, B.D. Behaviorism and positivism. Journal of the History of the Behavioral Sciences, 1972, 8, 222-231.

Mackenzie, B.D. Behaviorism and the limits of scientific method. Atlantic Highlands, N.J.: Humanities Press, 1977.

McKeachie, W.J. Psychology in America's bicentennial year. American Psychologist, 1976, 31, 819-833.

McNemar, Q. Lost: our intelligence? Why? American Psychologist, 1964, 19, 871-882.

McNemar, Q. Psychological statistics (4th ed.). New York: Wiley, 1969.

Pearlman, K., Schmidt, F.L., & Hunter, J.E. Validity gen-

eralization results for tests used to predict job-proficiency and training success in clerical occupations. Journal of Applied Psychology, 1980, 65, 373-406.

Prien, E.P., & Ronan, W.W. Job analysis: A review of research findings. Personnel Psychology, 1971, 24, 371-396.

Schmid, J., & Leiman, J.M. The development of hierarchical factor solutions. Psychometrika, 1957, 22, 53-61.

Schmidt, F.L. Poor hiring decisions lower productivity. Civil Service Journal, Jan./Mar., 1979, p.9.

Schmidt, F.L., & Hunter, J.E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

Schmidt, F.L., & Hunter, J.E. Moderator research and the law of small numbers. Personnel Psychology, 1978, 31, 215-232.

Schmidt, F.L., & Hunter, J.E. New research findings in personnel selection: Myths meet realities in the 1980's. Schmidt, F.L., & Hunter, J.E. Employment testing: Old theories and new research findings. American Psychologist, 1981, 36, 1128-1137.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. Task differences as moderators of test aptitude validity in

selection: A red herring. Journal of Applied Psychology, 1981, 66, 166-185.

Schmidt, F.L., Hunter, J.E., Pearlman, K., & Shane, G.S. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 1979, 32, 257-281.

Schmidt, F.L., Pearlman, K., & Hunter, J.E. The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. Personnel Psychology, 1980, 33, 705-724.

Schull, W.J., & Neel, J.V. The effects of inbreeding on Japanese children. New York: Harper & Row, 1965.

Spearman, C. The abilities of man. New York: Macmillan, 1927.

Wonderlic, E.F., & Wonderlic, C.F. Wonderlic Personnel Test: Negro norms. Northfield, Illinois: E.F. Wonderlic & Associates, Inc., 1972.

Footnotes

This article is based on an invited address (Division 14, Industrial and Organizational Psychology) presented at the convention of the American Psychological Association in Los Angeles, California, August 26, 1981. The author is indebted to Frank L. Schmidt and Kenneth Pearlman for their constructive criticisms of an earlier draft of this article.

Requests for reprints should be sent to Arthur R. Jensen, Institute of Human learning, University of California, Berkeley, California 94720.

[1]Assuming an average sample size of 150 and nine predictor variables (as there are nine subtests in the GATB), the observed multiple correlation of +.36, when corrected for bias, using the usual shrinkage formula (McNemar, 1969, p. 205) becomes +.27.

Figure Captions

Figure 1. Frequency distribution of 537 validity coefficients of the General Aptitude Test Battery for 446 different occupations. G score is general intelligence; multifactor validity is based on an optimally weighted composite of nine GATB aptitudes (including G) for each job category. The median validities are +.27 for G and +.36 for the multifactor composite.

Figure 2. The correlation (Pearson r) between choice RT and IQ as a function of the alternatives (n), in a group of 48 Ss with Wechsler Performance IQs ranging from 57 to 130. (From Lally & Nettelbeck, 1977.)

Figure 3. The g factor loadings of eleven WISC subtests plotted as a function of the percent of inbreeding depression on subtest scores. (Information, Comprehension, Arithmetic, Similarities, Vocabulary, Picture Completion, Picture Arrangement, Block Design, Object Assembly, Digit Symbol, Mazes.)

Figure 4. Profiles of (a) the correlations of each of the nine GATB aptitudes with IQ, (b) the g factor loadings of each aptitude, and (c) the mean white-black difference (in σ units) of each of the aptitudes.

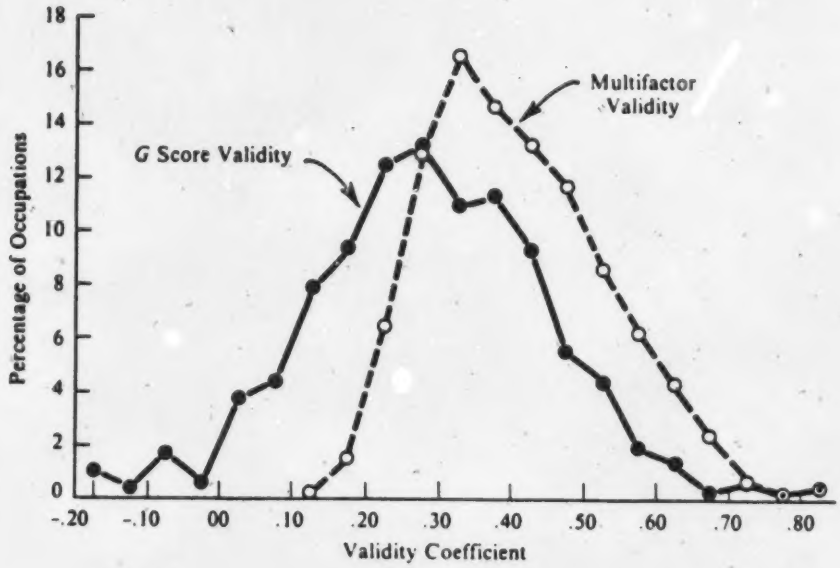

Figure 1. Frequency distribution of 537 validity coefficients of the General Aptitude Test Battery for 446 different occupations. G score is general intelligence; multifactor validity is based on an optimally weighted composite of nine GATB aptitudes (including G) for each job category. The median validities are +.27 for G and +.36 for the multifactor composite.

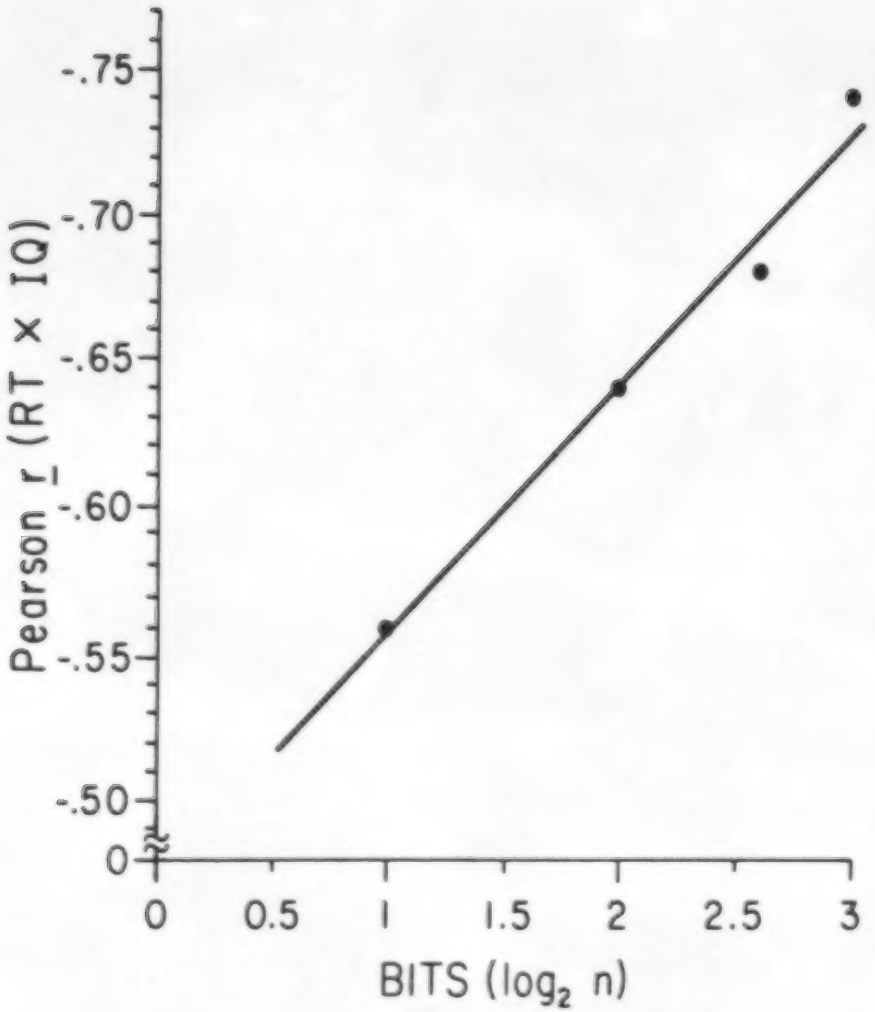

Figure 2. The correlation (Pearson r) between choice RT and IQ as a function of number of alternatives (n), in a group of 48 Ss with Wechsler Performance IQs ranging from 57 to 130. (From Lally & Nettelbeck, 1977.)

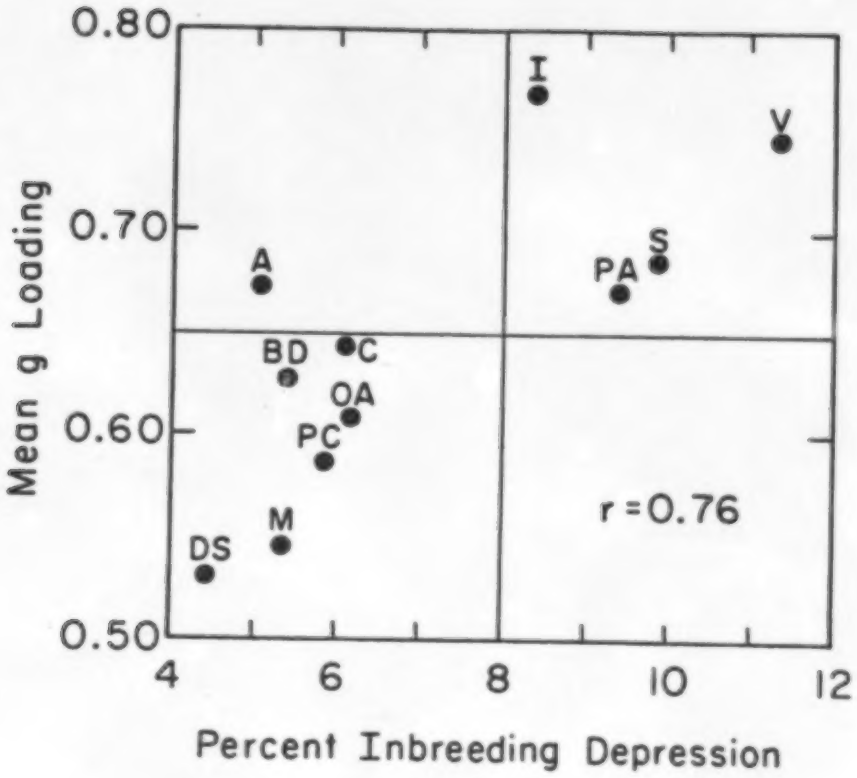

Figure 3. The g factor loadings of eleven WISC subtests plotted as a function of the percent of inbreeding depression on subtest scores. (Information, Comprehension, Arithmetic Similarities, Vocabulary, Picture Completion, Picture Arrangement, Block Design, Object Assembly, Digit Symbol, Mazes.)

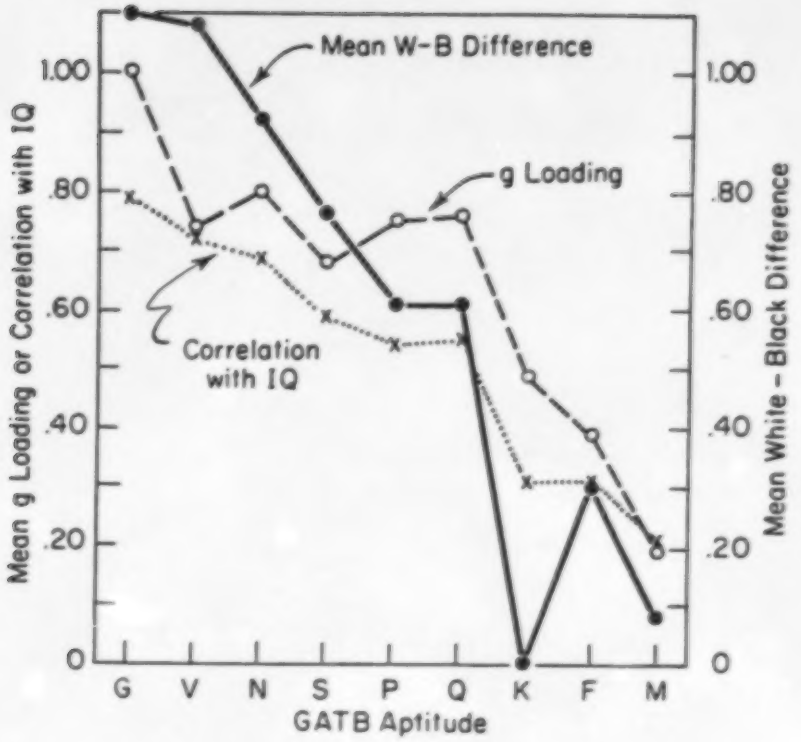

Figure 4. Profiles of (a) the correlations of each of the nine GATB aptitudes with IQ, (b) the g factor loadings of each aptitude, and (c) the mean white-black difference (in σ units) of each of the aptitudes.